NSTL 集成第三方来源元数据的实践与探索

于倩倩[1,2]　张建勇[1,2]

（1 中国科学院文献情报中心；2 NSTL 数据研究管理中心）

摘要：本文在分析 Web of Science、Scopus 和 NSTL 联合数据加工系统采用的文献元数据规范基础上，以期刊论文为例，实现三者数据库元数据的映射，并对元数据的描述方式和特点进行了比较，同时结合实践，提出了元数据映射过程中需要注意的问题，为相关系统的元数据建设和利用第三方数据提供参考依据。

关键词：Web of Science, Scopus, NSTL，元数据

1 引言

国家科技图书文献中心（NSTL）"十三五"发展规划提出，要优化国家科技文献资源保障体系，拓展元数据资源采集方式。通过与国内外出版商、相关信息机构协商、购买、交换、赠与、缴存等多渠道获取第三方元数据，是拓展建设元数据资源的方式之一。因此，需要在 NSTL 联合数据加工系统采用的文献资源元数据规范基础上，深入分析其它来源数据的类型特点和建设需求，建立健全元数据规范，以便有效地集成第三方来源元数据。

然而，不同的文献数据库，元数据规范描述的内容和方式往往存在差异，这影响了相互间的数据交互和共享，也对集成和利用其它数据库资源产生障碍。第三方元数据格式的多样性与 NSTL 加工资源需求接口单一性之间的冲突，使得第三方来源元数据与 NSTL 文献资源元数据之间的互操作成为必然。明确元数据的内容和组织方式，制订相关规则实现第三方来源与 NSTL 文献资源元数据的映射，是有效集成和利用第三方数据库数据的可操作方式之一。

本文在分析 Web of Science（以下简称 WOS）数据库[1]、Scopus 数据库[2]和 NSTL 采用的文献资源加工规范[3]基础上，结合相关实践，以期刊论文为例，对三者的元数据映射内容、映射效果、元数据描述方式和特点进行比较，以期为相关文献信息系统的元数据建设和利用已有第三方数据库资源提供借鉴。

2 期刊论文元数据结构

根据 DC 元数据设计的模块化原则[4]，并结合分析三个文献数据库的元数据内容，期刊论文元数据可以分为论文元数据、作者元数据、作者机构元数据、期刊元数据、会议元数据、基金元数据、参考文献元数据、施引文献元数据等。按照实体分析法，期刊论文实体间的关系如图 1 所示，一篇期刊论文可能由一个或多个作者撰写，一个作者属于一个或多个机构，论文发表在期刊上，可能来自于某个会议，也可能挂靠某个基金，可能具有 1 篇或多篇参考文献，也可能被 1

篇或多篇文献引用等。

WOS、Scopus、NSTL 的期刊论文元数据类型如表 1 所示，可以看出，WOS、Scopus 对 8 类元数据均有描述，NSTL 缺乏对会议、基金和施引文献元数据的描述。分析原因，一是，WOS、Scopus 使用一套元数据 Schema 描述多种文献类型如期刊论文、会议论文、图书、专利等，因此，如果期刊论文中涉及到会议、基金信息，会出现相关的会议、基金描述信息，NSTL 虽然相同含义字段元素通用，但以文献类型为基础划分元数据 Schema，会议元数据出现在会议论文 Schema 中；二是，NSTL 目前还没有基金数据、施引文献数据的描述。

表 1 WOS、Scopus、NSTL 的期刊论文元数据

元数据类型	论文	作者	作者机构	期刊	会议	基金	参考文献	施引文献
WOS	√	√	√	√	√	√	√	√
Scopus	√	√	√	√	√	√	√	√
NSTL	√	√	√	√			√	

3 元数据映射与比较

以 NSTL 期刊论文元数据（部分字段是必备(Required)字段，以 R 表示）为基础，对比 WOS、Scopus 在论文元数据、作者/机构元数据、期刊元数据、参考文献元数据中相同字段的描述内容和方式，并分析不同文献数据库元数据描述的特点，以期取长补短，更好地对文献数据进行管理。

3.1 论文元数据的映射比较

NSTL 论文描述信息是期刊论文描述元数据规范的主体部分，描述的内容包括论文题名、关键词、文摘和分类信息等几个部分。WOS 论文描述信息包括论文唯一标识 UID、题名、文献类型信息等，起页、止页、总页数在期刊元数据中进行描述。Scopus 论文描述信息包括论文题名、摘要、文献类型等信息，主题词、分类号信息在增强描述集中描述，起页、止页、总页数在期刊元数据中描述，参考文献总数在参考文献元数据中描述，论文唯一标识符包括 eid、pui、pii 等。WOS、Scopus 与 NSTL 论文元数据映射如表 2 所示。

表 2 论文元数据映射

元数据标签	NSTL Schema	WOS Schema	Scopus Schema
记录号	paper_id(R)		
题名	title (R)	title type=”item”	titletext original="y"
其他语种题名	alternative	title type=”foreign”	titletext original="n"
文摘	abstract	abstract_text	abstract original="y"
其它语种文摘	abstract_alternative		abstract original="n"

关键词	keyword	keyword	author-keyword
其它语种关键词	keyword_alternative		
主题词	subject_heading	subject	mainterm
主题词表	thesaurus		descriptors controlled="y" type=””
分类号	classification		classification
分类法	classification_scheme		classifications type=””
正文语种	language (R)	language	citation-language xml:lang=””
其它语种	other_language		
起页	start_page (R)	page begin=””	pagerange first=””
止页	end_page	page end=””	pagerange last=””
总页数	total_page_number (R)	page page_count=””	pagecount
参考文献总数	total_reference_number	refs count=””	refcount=””
文献号	paper_no		
本地唯一标识符	local_doi (R)		
DOI	doi	identifier type="doi" value=””	doi
论文类型	paper_type		
资源类型	type(R)	doctype	citation-type code=””

从表 2 可以看出，在 22 个 NSTL 论文元数据字段中，WOS 有 11 个字段实现映射，仅占 1/2，Scopus 有 16 个字段完成映射。在未映射的字段中还包含了必备字段 paper_id 和 local_doi，这样的话，如果想要将映射后的数据以 NSTL Schema 格式输出则无法完成，解决方法可以将必备字段取值为内部 id 或输出为空标签。

另外，不同数据库相同字段的元数据枚举值可能不同，例如 NSTL 的 type、WOS 的 doctype、Scopus 的 citation-type，虽然都是描述文献的类型，但三者的类型值不同，需要进行统一或指定枚举值映射方式；同一字段的元素取值类型可能不同，例如 NSTL 中作者顺序 author_sequence 取值类型为 byte[5]，WOS 中 seq_no 取值类型为 positiveInteger，需要调整为一致。

从表 2 中还可以看出，NSTL 通过元素方式进行描述，WOS、Scopus 多用属性进行描述，例如，在 WOS Schema 中，描述题名(title)元素的属性有类型(type)，type 的取值除了论文(item)、其它语种(foreign)还包含出版物(source)、iso 出版物缩写(abbrev_iso)、11 位出版物缩写(abbrev_11)等，页码、参考文献数等都采用了属性限定元素的方式，更好地将描述内容进行归并。

3.2 作者/机构元数据的映射比较

在 NSTL 中，作者是指期刊论文撰写者，在 WOS、Scopus 中，论文作者与出版者、图表制作者、翻译者等共用子元素，因此需要指定父元素 author 才能实

现准确映射，如表 3 所示。除了映射元素外，WOS、Scopus 中都有对作者姓、名、通讯作者、机构地址、所属国家和城市的描述，以及唯一标识符的描述。WOS 中作者的唯一标识符包括 ResearcherID、ORCID、dais_id 等，Scopus 中作者唯一标识符为 AuthorId，机构唯一标识符为 afid，均是可选属性。作者唯一标识符对唯一识别作者具有重要作用。

表 3 作者/机构元数据映射

元数据标签	NSTL Schema	WOS Schema	Scopus Schema
作者顺序	author_sequence (R)	name seq_no=”” 且 role=”author”	author seq=””
作者姓名	author_name(R)	name(role=”author”) display_name	author indexed-name
其它形式作者姓名	author_name_alternative	有 full_name 时对照 name(role=”author”) wos_standard; 无 full_name，无对照字段	author initials
作者所属机构	affiliation	address_name address_spec organization	affiliation organization
其它形式机构	affiliation_alternative		
作者 Email 地址	email	name(role=”author”) email_addr	author e-address type="email"

在 WOS 数据库中，通过 addr_no 建立了作者和机构之间的一一对应关系，如果作者姓名(name)元素中的属性 addr_no 和地址 address_spec 元素中的属性 addr_no 相同，则表示此机构是该作者的机构。这样，不管作者有几个机构，都可以方便地实现对应，避免重复记录。

相较于 WOS 对作者和机构信息的对应描述，NSTL 和 Scopus 的表达方式相对繁琐。NSTL 顺序描述作者和机构信息，如果文献作者隶属于同一机构，则会出现多次相同机构的描述信息，造成机构信息的冗余。Scopus 以机构为基准对作者进行划分，同一机构的作者会出现在同一描述记录中，如果作者属于多个机构，则会在多个描述记录中出现该作者的姓名和联系方式等描述信息，造成作者信息的冗余。

3.3 期刊元数据的映射比较

期刊是期刊论文的载体，在 NSTL 中，期刊元数据包括期刊描述信息见表 4 中的前 14 个元素和卷期描述信息见表 4 中的后 3 个元素，在 WOS、Scopus 中卷期描述信息包含在期刊描述信息中。除了表 4 中的映射元素，WOS 包含了更详细的期刊名称的缩写信息、卷期出版日期信息和出版商地址信息，Scopus 描述了期刊唯一标识符 srcid、期刊名称缩写、文献来源网址、期刊编辑者信息等。

表 4 期刊元数据映射

元数据标签	NSTL Schema	WOS Schema	Scopus Schema
母体数据源编号	catalog_code(R)		
订购号	subscription_number		
ISSN	issn	identifier type="issn"	issn type="print"
EISSN	eissn	identifier type="eissn"	issn type="electronic"
CODEN	coden		codencode
国内统一书刊号	cn	identifier type="cn"	
母体文献名称	host_title(R)	title type="source"	sourcetitle
其他语种母体文献名称	host_title_alternative		translated-sourcetitle
语种	host_language(R)		
母体文献分类号	host_classification		
出版地	publishing_place	publisher address_spec city	publisher affliation city
出版者	publisher	publisher name (role="publisher")display_name	publishername
起始年	start_year(R)		
终止年	end_year		
卷期出版年	year(R)	pub_info pubyear=""	publicationyear first=""
卷信息	volume	pub_info vol=""	voliss volume=""
期信息	issue	pub_info issue="" part_no="" supplement="" special_issue=""	voliss issue="" or supplement

在 NSTL17 个元数据字段中，WOS 有 9 个字段实现映射，Scopus 有 10 个字段实现映射，对于未实现映射的必备字段处理方式同论文元数据未映射的必备字段。在 NSTL 中，只有期信息字段，没有划分增刊、特刊、分期字段，但指定了这些字段在期信息字段中的著录规则，例如有期号，但该期又分为若干分期的，分期前缀照录，增刊、专刊填写在期号后，若无期号则直接填写增刊信息等[6]，可根据这些著录规则对 WOS、Scopus 相应数据进行数据抽取合并。

3.4 参考文献元数据的映射比较

在 NSTL 中，参考文献内容包括引文作者、题名、出处、卷期以及获取访问路径等。参考文献信息可以让用户从作者研究脉络角度查找到一组相关文献[7]。WOS 包含了参考文献中的作者、题名、刊名、卷、页信息，没有参考文献原始信息字段，Scopus 既包含了原始信息字段，也包含了作者、题名等拆分字段。

三者参考文献元数据映射如表 5 所示，对于未实现映射的必备字段处理方式同前。

表 5 参考文献元数据映射

元数据标签	NSTL Schema	WOS Schema	Scopus Schema
引文类型	citation_type(R)		
引文原始信息	citation_orig_info(R)		ref-fulltext
引文第一作者	citation_author1	citedAuthor	ref-authors author seq=”1”
引文第二作者	citation_author2		ref-authors author seq=”2”
引文第三作者	citation_author3		ref-authors author seq=”3”
引文题名	citation_title	citedTitle	ref-titletext
引文出处	citation_sourcetitle	citedWork	ref-sourcetitle
引文出版年	citation_year	reference year	ref-publicationyear first=””
引文卷号	citation_volume	reference volume	ref-volisspag voliss volume=””
引文期号	citation_issue		ref-volisspag voliss issue=””
引文页	citation_page	reference page	ref-volisspag pagerange first="" last=""
引文主编	citation_editor_in_chief		
引文出版者	citation_publisher		
链接地址	citation_url		

在 WOS 中，通过一篇论文作为参考文献的次数可以得到此论文的被引次数。原文献中具有唯一标识符 UID，参考文献也包含唯一标识符 ut，如果 UID 与 ut 值相同，则表示这是同一篇文献，具有 UID 的文献被包含 ut 的文献引用了，UID 与多少个 ut 值匹配上，就表示具有此 UID 的文献被引用了多少次。需要注意的是，文献的 UID 是不会变的，但参考文献的 ut 值可能会因为施引文献的改动而被删除或被替代，此外有些参考文献的 ut 值无法获得、数据更新或修改等都可能影响文献被引次数的计算。

4 元数据映射方式的优势和不足

对于同一篇期刊论文，不同的文献数据库，信息组织方式和内容揭示方式都存在差异，有的数据库描述信息更加详细，元数据字段更加齐全，有的数据库描述信息相对简略，元数据字段也相对较少，而且不同的数据库元数据侧重点也有所不同。在不同数据库间元数据较难实现统一的情况下，通过元数据映射的方式，是实现数据集成管理的可行方式。对于不同的数据库，元数据字段映射的数量越多，外部来源数据库的数据利用越充分。

通过对 WOS、Scopus 与 NSTL 元数据的映射，可以看出，WOS、Scopus 与 NSTL 在通用字段上描述相同，能够进行映射，但未实现全部字段的映射，一定程度上影响了加工数据的全面性。另外，WOS、Scopus 对作者、机构、期刊等有更多较为详细的描述字段，在 NSTL 中没有体现，这些数据对于文献资源信

息的揭示更为细颗粒化，通过元数据映射输出的方式，也在一定程度上造成了这些外部数据源数据的失真。

此外，WOS、Scopus 在不同的元数据描述粒度上设置了多种唯一标识，例如 WOS 对论文(uid)、作者(r_id)、期刊卷期(ids)等有唯一标识的描述，Scopus 对论文(eid)、作者(auid)、机构(afid)、期刊(srcid)等有唯一标识的描述。在 NSTL Schema 中添加外部论文唯一标识字段，与其它数据库唯一标识进行映射，可以唯一识别来自于 WOS 等外部数据库的论文，添加外部作者、作者机构、期刊等的唯一标识还可以对这些数据进行唯一识别，区分自加工数据与外部来源数据。

5 结语

在当前不同文献数据库元数据描述字段不尽相同的情况下，如果相互之间的元数据能够进行映射，对实现不同数据库之间数据的交互和流转具有重要意义，元数据字段映射数量越多，数据越能得到充分利用。本文以 NSTL 期刊论文元数据为基础，完成了 WOS、Scopus 元数据与 NSTL 元数据的映射，分析了三者元数据描述的特点，并提出元数据映射过程中需要注意的问题。NSTL 目前已将购买的 WOS 数据装入准备库，陆续还将装入其它文献数据库数据，装入的数据根据不同数据库 Schema 对照表以 NSTL Schema 格式输出，将加工人员从繁琐的加工工作中解脱出来，并通过利用第三方数据库数据大大提升了 NSTL 数据加工速度和系统的自动化水平。

---

[1] Web of Science[EB/OL].[2014-05-08].www.webofknowledge.com/WOS.
[2] Scopus[EB/OL].[2014-06-18]. https://www.scopus.com/.
[3] 张建勇,曾燕.文献数据库数据加工规范[M].知识产权出版社,2009.
[4] The Singapore Framework for Dublin Core Application Profiles[EB/OL].[2015-05-08].http://dublincore.org/documents/singapore-framework/.
[5] NSTL_journalarticle.xsd[EB/OL].[2015-05-20]. http://spec.nstl.gov.cn/specification/namespace/NSTL_journalarticle.xsd
[6] Issue[EB/OL].[2015-05-22].http://spec.nstl.gov.cn/specification/index.php?title=Issue
[7] 期刊论文描述元数据规范[EB/OL].[2015-06-05]. http://spec.nstl.gov.cn/specification/index.php?oldid.